

# An analysis of rainfall based on entropy theory

Vicente de Paulo Rodrigues da Silva<sup>1\*</sup>; Adelgicio Farias Belo Filho<sup>1</sup>; Enio Pereira de Souza<sup>1</sup>, Célia Campos Braga<sup>1</sup>; Romildo Morant de Holanda<sup>2</sup>; Rafaela Silveira Rodrigues Almeida<sup>1</sup> and Armando César Rodrigues Braga

Federal University of Campina Grande, Av. Aprígio Veloso, 882, Bodocongó, 58109 970, Campina Grande, PB, Brazil

<sup>2</sup>Federal Rural University of Pernambuco, R. Manuel de Medeiros, S/N - Dois Irmãos, Recife - PE, 52171-900

<sup>3</sup>Federal University of Paraíba, Cidade Universitária, s/n - Castelo Branco, João Pessoa - PB, 58051-900

\* Corresponding author. Tel./fax: +55 8333101202.

E-mail address: vicente.paulo@ufcg.edu.br (V.P.R. Silva).

**Abstract**— *The principle of maximum entropy can provide consistent basis for analyzing rainfall and for geophysical processes in general. The daily rainfall data was assessed using the Shannon entropy for a 10-years period from 189 stations in the northeastern region of Brazil. Mean values of marginal entropy were computed for all observation stations and isoentropy maps were then constructed for delineating annual and seasonal characteristics of rainfall. The Mann-Kendall test was used to evaluate the long-term trend in marginal entropy for two sample stations. The marginal entropy values of rainfall were higher for locations and periods with highest amount of rainfall. The results also showed that the marginal entropy decreased exponentially with increasing coefficient of variation. The Shannon theory produced spatial patterns which led to a better understanding of rainfall characteristics throughout the northeastern region of Brazil. Trend analysis indicated that most time series did not have any significant trends.*

**Keywords**— *Mann-Kendall test, Information transfer, Measure the disorder.*

## I. INTRODUCTION

The concept of entropy was advanced later in the works of in quantum mechanics, and was reintroduced in information theory by Shannon (1948) as a measure of information, disorder or uncertainty. The Shannon entropy has since been employed in numerous areas (Singh and Rajagopal, 1987), such as mathematics (Dragomir et al., 2000), economics (Kaberger and Mansson, 2001), ecology (Ricotta, 2002), climatology (Kawachi et al., 2001), medicine (Montaño et al., 2001) and hydrology (Singh, 1997). One measure of uncertainty or disorder of a variable is entropy. Entropy can be calculated if the probability distribution function or probability density

function of the random variable is given in a discrete or continuous form, using the informational entropy theory.

An interesting application of entropy has been for reducing the gap between information needs and data collected by monitoring networks (Krstanovic and Singh, 1993a; Krstanovic and Singh, 1993b; Al-Zahrani and Husain, 1998; Agrawal et al., 2005; Chen et al., 2007). In this application, stations are evaluated by transmission of information to and from stations (Markus et al. 2003). Likewise, entropy has been used for assessing the space variability of rainfall, one of the primary constraints to water resources development and water use practices (Silva et al., 2003; Mishra et al., 2009). The main point here is to measure the disorder or uncertainty of the occurrence of rainfall by entropy (Maruyama et al. 2005).

Chen et al. (2007) suggested that the variability of rainfall can be more appropriately measured by the Shannon entropy and hence rainfall characteristics of 1-day resolution time series can be described. Thus, the entropy theory, comprising the Shannon entropy, seems to have much potential that remains yet to be fully exploited. Most of works have mainly focused on the spatial and temporal variability of rainfall using information theory for temperate zones while less attention has been given to methodologies that include rainfall in tropical climate zones for improving the estimates of rainfall variability at a time scale from years to days by exploiting the time series structure. The identity of the cumulative sources of uncertainty in rainfall remains practically unknown and have not yet been investigated in systematic manner. To address this issue, we used the Shannon entropy to quantify the variability of rainfall in the northeastern region of Brazil and assess long-term trends in marginal entropy of annual and seasonal rainfall using the Mann-Kendall test.

## II. MATERIAL AND METHODS

### Shannon entropy

The discrete form of the Shannon entropy was obtained by (Kawachi et al., 2001):

$$H(p_1, p_2, \dots, p_k) = -k \sum_{i=1}^{i=n} p_i \ln p_i \quad (1)$$

where  $p_i$  is a probability of the  $i$ th outcome of a discrete random variable,  $n$  is the number of events and  $k$  is a positive constant, which depends on the choice of measurement units. For all same  $p_i$ s the entropy is  $H = \log_2 n$ , which is a monotonically increasing function of  $n$ . The units of entropy depend on the base of the logarithm in Eq. (1), and are bits (binary digits) for base 2, and napiers or nats for base  $e$ , while the term Hartley has been proposed for base 10. Taking  $k = 1$  and the base of the logarithm as 2, bit is used as a unit of measurements of entropy. Entropy  $H(p_i)$  is also called marginal entropy of a univariate variable  $p$ .

The annual rainfall ( $R$ ) for each hydrologic event can be obtained by:

$$R = \sum_{i=1}^{i=365} r_i \quad (2)$$

where  $r_i$  is daily value of rainfall for the  $i$ th day of the year. The occurrence probability of the rainfall amount on the  $i$ th day was expressed as the relative frequency ( $p_i$ ) as:

$$p_i = \frac{r_i}{R} \quad (3)$$

### Mann-Kendall test

The Mann-Kendall non-parametric test (Mann 1945; Kendall 1975) was applied for assessing the trend of rainfall time series. This test is based on statistic  $S$  defined as:

$$S = \sum_{i=2}^n \sum_{j=1}^{i-1} \text{sign}(x_i - x_j), \quad (4)$$

where  $x_j$  are the sequential data values,  $n$  is the length of the time series and  $\text{sign}(x_i - x_j)$  is -1 for  $(x_i - x_j) < 0$ , 0 for  $(x_i - x_j) = 0$ , and 1 for  $(x_i - x_j) > 0$ . The mean  $E[S]$  and variance  $V[S]$  of statistic  $S$  may be given as:

$$E[S] = 0 \quad (5)$$

$$\text{Var}[S] = \frac{n(n-1)(2n+5) - \sum_{p=1}^q t_p(t_p-1)(2t_p+5)}{18} \quad (6)$$

where  $t_p$  is the number of ties for the  $p$ th value and  $q$  is the number of tied values. The second term represents an

adjustment for tied or censored data. The standardized test statistic ( $Z_{MK}$ ) is computed as:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & \text{if } S < 0 \end{cases} \quad (7)$$

The presence of a statistically significant trend was evaluated for testing the null hypothesis that no trend existed. A positive  $Z_{MK}$  value indicates an increasing trend while a negative one indicates a decreasing trend. To test for either increasing or decreasing monotonic trend at  $p$  significance level, the null hypothesis was rejected if the absolute value of  $Z_{MK}$  was greater than  $Z_{1-p/2}$ , which was obtained from the standard normal cumulative distribution table. In this study, the significance levels of  $p = 0.01$  and  $0.05$  were applied. The non-parametric estimate of the magnitude of the slope of trend was obtained as follows (Hirsch et al. 1982).

$$\beta = \text{Median} \left[ \frac{(x_j - x_i)}{(j - i)} \right] \quad \text{for all } i < j \quad (8)$$

where  $x_j$  and  $x_i$  are the data points measured at times  $j$  and  $i$ , respectively.

### Study area

The northeastern region of Brazil, bounded to the north and east by the Atlantic Ocean, covers an area of about 1.5 million square kilometers. Approximately 60% of this region is a semi-arid area. The area is inhabited by more than 30 million people and the economy is mainly based on subsistence rainfed crop production. The northeastern region is influenced by several large-scale precipitation mechanisms. The rainy-season occurs between January and June and the dry-season between July and December. The wet-season occurs between March and May and the normal annual rainfall ranges from 400 to 2000 mm (Silva, 2004). The region is dominated by semi-arid climate with heterogeneous vegetation cover and the mean air temperature varies between 15 and 33 °C (Silva et al., 2006).

The temporal trend in the entropy time series was analyzed using data from two weather stations. These stations are located in the state of Ceará, namely Icô (latitude: 6°24'04" S; longitude: 38°51'44" W; altitude: 153.4 m above sea level) and São Luiz do Curu (latitude: 3°40'12" S; longitude: 39°14'36" W; altitude: 38.4 m above

sea level). The isoinformation contours are then drawn showing areas of greater or less information transfer. The results of the study are compared to the variance approach. Variance is a measure of dispersion and its simplicity remains its major attraction. Variance has had a primordial role in the analysis of variability (Mishra et al., 2009)

#### Rainfall data

Daily time series of rainfall recorded at 189 stations for a minimum period of 10 years in the northeastern region of Brazil were analyzed and annual totals of marginal entropy were obtained. The mean entropy values computed for observation stations were employed to construct the isoentropy maps in order to delineate rainfall characteristics. Annual period and rainy and dry season rainfall time series were used to assess long-term trends in marginal entropy and their coefficients of variation from two sample stations: Icó station for a period of 1957 to 2001 and São Luiz do Curu station for a period of 1968 to 2001.

### III. RESULTS AND DISCUSSION

Mean values of marginal entropy and the coefficient of variation (CV) of rainfall at two sample stations in northeastern region of Brazil for the annual and dry and rainy seasons are shown in Table 1. For both stations, marginal entropy values of rainfall were low during the dry season and high during the rainy season. The values of mean annual entropy were very similar to those for the rainy season, when the total rainfall during the dry season was comparatively smaller than that of rainy season. The variability of annual time series has higher disorder in comparison to constituent seasonal time series. Different seasons contribute differently to the variability of annual rainfall time series. The rainy season variability contributes more to the variability of annual time series, whereas dry season contributes less to the annual variability.

For Icó station, 86% of the annual entropy of rainfall was observed in the rainy season. Similarly, for the São Luiz do Curu station, 87% of the annual entropy of rainfall was observed in the rainy season. In general, the coefficient of variation was high. The CV values of the marginal entropy reached a maximum of 114.7% at the São Luiz do Curu station for rainy season rainfall and a minimum of 34.2% at the Icó station for annual rainfall. The most common statistic used to describe variability is variance, which measures the spread in a data set. However, the variability of rainfall time series can be quantitatively measured by using entropy which can be described in spatial and temporal terms (Mishra et al.

(2009). The opinions are conflicting between variance and entropy for analyzing variability in times series. For example, Soofi (1997) considers that the interpretation of variance as a measure of uncertainty must be done with caution. However, according to Maasoumi (1993), entropy can be an alternative measure of dispersion. According to Ebrahimi et al. (1999), both these measures reflect concentration but their metrics for concentration are different.

The coefficients of determination of 0.95 and 0.99 were obtained in Luiz do Curu station and Icó station, respectively (Figure 1). As expected, a good relationship is evident because marginal entropy is also a variability measure of time series. Silva et al. (2003) who assessed the evaluation of the rainfall variability in Paraíba state, Brazil, using entropy theory showed that for any time series the entropy decreases exponentially with increase of standard deviation. Our results also showed that there was no an indefinite exponential increase of marginal entropy for rainfall since such increase occurred until it reaches the maximum entropy. This is consistent with the second law of thermodynamics which states that the entropy of an isolated system tends to increase until it reaches equilibrium. In this context, Ebrahimi et al. (1999) examined the role of variance and entropy in ordering distributions and random prospects, and concluded that there is no universal relationship between these measures in terms of ordering distributions.

Annual and seasonal values of marginal entropy of rainfall for São Luiz do Curu and Icó stations are shown in Fig. 2. Despite decreasing trends in annual and rainy season rainfall at both stations (Table 2), an increasing trend of marginal entropy in rainfall was observed during the year and rainy and dry seasons. Trend analysis indicated that most time series did not have any significant trends. Although Shannon entropy is a quantification of the amount of information within a dataset, its static probabilistic nature cannot capture the temporal variability of information. It therefore shows no sensitivity in time. Results support the theoretical observations that Shannon entropy is strongly related to the CV relationship, and it is suggested that this is likely to provide a more robust measure of variability than those in CV. This issue is particularly relevant because entropy is insensitive to timing errors. This makes it dangerous as a stand-alone measure, but potentially provides a useful diagnostic in spatial variability. Rainfall data presented an increasing trend during the dry season at Icó station, but the time series was not statistically significant based on the Mann-Kendall test. These results suggested that the temporal trend of entropy was not influenced by the original data.

On this issue, Kawachi et al. (2001) showed that average annual entropy and average annual rainfall were less mutually related with a coefficient of correlation of 0.19. Our results evidence that the trend in marginal entropy was statistically significant for annual rainfall at São Luiz do Curu station based on the Mann-Kendall test ( $p < 0.05$ ) and for dry season ( $p < 0.01$ ).

Spatial distributions of isoentropy in annual and rainy and dry rainfall in the northeastern region of Brazil are shown in Fig. 3. The isoentropy lines of marginal entropy of annual and dry season rainfall were higher throughout coast east of the region (Figs. 3A, 3C). However, higher values of isoentropy during the rainy season were located in the northern part of the region (Fig. 3B). As a natural consequence, higher rain might occur alternately during other periods of the year over the northeastern region. Minimum and maximum values of isoentropy in rainfall are observed in the same area for all analyzed periods. For instance, marginal entropies values of annual rainfall were minimum in the central area of northeastern region of Brazil, which corresponded to most of the semi-arid region.

The entropy values of rainfall were maximum in eastern and northern areas of the region which corresponded to most northeastern rainy areas. During the rainy season, the entropy decreased from 5.5 in the north to 1.5 bits in the south for rainfall. On the other hand, during the dry season entropy values of rainfall reached minimum values as compared to the other two periods as a consequence of the rainfall reduction. Mishra et al. (2009) also used marginal entropy to investigate the temporal variability of rainfall time series for the State of Texas, USA. They observed distinct spatial patterns in annual series and different seasons and that the variability of rainfall amount as well as number of rainy days within a year increased from east to west of Texas. The spatial distribution of marginal entropy was practically uniform during the dry season over almost the entire region, particularly for rainfall, with a mean value about of 0.5-1.5 bits. Martín and Rey (2000) analyzed the role of entropy to provide some mathematical arguments for justifying the use and interpretation of entropy as a measure of diversity and homogeneity.

As shown in Fig. 3, the isoentropy lines of rainfall divided the whole study region into two clusters, at left with higher values in entropy and at right with lower values of entropy. The marginal entropy of rainfall was high in areas and periods with the highest amount of rainfall. Results also demonstrate that the rainfall variability is higher in the semi-arid areas than in coast areas of the northeastern region. This indicates the

availability of water resources is low and should be used within the constraints. To meet the perennial water demands, proper planning is to be made to reduce the wastage of water as well as to store the excess water during time of precipitation. When performed a study to assess the stream gaging network in the State of Illinois, USA, Markus et al., (2003) showed that the correlation coefficient between entropy and least square regression method are inversely proportional to the information transmitted. Besides, stations located in an area of high gage density tend to receive and transmit more information. Inversely, gages having less significant regional value transmit substantially less information than they receive.

Despite large variations of marginal entropy for rainfall between periods and even between stations, the overall analysis showed much less variation of entropy during the dry season. Rainfall constitutes the primary input to the hydrologic cycle, and can thus be perceived to represent the potential water resources availability of an area. The disorder or uncertainty in the intensity and occurrence of rainfall in time is one of the primary constraints to water resources development and the water use practices. Distinct spatial patterns in annual series and different seasons were observed. For the three analyzed periods the entropy decreased from South to North. The results also indicated that highly disorderliness in the amount of rainfall during rainy season.

#### IV. CONCLUSIONS

The entropy concept was used in this study to determine the spatio-temporal variability/disorder of rainfall in northeastern region of Brazil. Entropy leads to a better understanding of time and space structure rainfall in the study area. It is shown that entropy can be effectively used for assessing the rainfall variability in both in space and time. The rainfall variability could satisfactorily be obtained in terms of marginal entropy as a comprehensive measure of the regional uncertainty of these hydrological events. The coefficient of variation is exponentially related to marginal entropy of rainfall, with the coefficient of determination close to 1. The Mann-Kendall test suggests that the temporal trend of entropy in rainfall is not influenced by the eventual trend of the original data.

#### REFERENCES

- [1] Agrawal D, Singh JK, Kumar A. 2005. Maximum entropy-based conditional probability distribution runoff model. *Biosystems Engineering* **90**: 103-113.
- [2] Al-Zahrani M, Husain T. 1998. An algorithm for designing a precipitation network in the south-

- western region of Saudi Arabia. *Journal of Hydrology* **205**: 205-216.
- [3] Chen YC, Wei C, Yeh HC. 2007. Rainfall network design using kriging and entropy. *Hydrological Process* **1**: 340-346.
- [4] Dragomir SS, Scholz ML A, Sunde J. 2000. Some upper bounds for relative entropy and applications. *International Journal of Computational and Applied Mathematics* **39**: 91-100.
- [5] Ebrahimi N, Maasoumi E, Soofi E. 1999. Ordering univariate distributions by entropy and variance. *Journal of Econometrics* **90**: 317-336.
- [6] Hirsch RM, Slack JR, Smith RA. 1982. Techniques of trend analysis for monthly water quality data. *Water Resources Research* **18**: 107-121.
- [7] Kaberger T, Mansson B. 2001. Entropy and economic processes – physics perspectives. *Ecological Economics* **36**: 165-179.
- [8] Kawachi T, Maruyama T, Singh VP. 2001. Rainfall entropy delineation of water resources zones in Japan. *Journal of Hydrology* **246**: 36-44.
- [9] Kendall MG. 1975. Rank correlation measures. London: Charles Griffin. 220.
- [10] Krstanovic PF, Singh VP. 1993a. Evaluation of rainfall networks using entropy: I. Theoretical development. *Water Resources Management* **6**: 279-293.
- [11] Krstanovic PF, Singh VP. 1993b. Evaluation of rainfall networks using entropy: II. Application. *Water Resources Management* **6**: 295-314.
- [12] Maasoumi EA. 1993. Compendium to information theory in economics and econometrics. *Econometric Reviews* **12**: 137-181.
- [13] Mann HB. 1945. Nonparametric tests against trend. *Econometrica* **13**: 245-259.
- [14] Markus M, Knapp HV, Tasker GD. 2003. Entropy and generalized least square methods in assessment of the regional value of streamgages. *Journal of Hydrology* **283**: 107-121.
- [15] Martín AA, Rey JM. 2000. On the role of Shannon's entropy as a measure of heterogeneity. *Geoderma* **98**: 1-3.
- [16] Maruyama T, Kawachi MT, Singh VP. 2005. Entropy-based assessment and clustering of potential water resources availability. *Journal of Hydrology* **309**: 104-113.
- [17] Mishra AK, Ozger M, Singh VP. 2009. An entropy-based investigation into the variability of precipitation. *Journal of Hydrology* **370**: 139-154.
- [18] Montaña MAJ, Ebeling W, Pohl T, Rapp PE. 2001. Entropy and complexity of finite sequences as fluctuating quantities. *BioSystems* **64**: 23-32.
- [19] Ricotta C. 2002. Bridging the gap between ecological diversity indices and measures of biodiversity with Shannon's entropy: comment to Izák and Papp. *Ecological Model* **152**: 1-3, 2002.
- [20] Shannon CE. 1948. The mathematical theory of communications. *Bell System Technical Journal* **27**: 379-423.
- [21] Silva VPR, Cavalcanti EP, Nascimento MG, Campos JHBC. 2003. Análises da precipitação pluvial no Estado da Paraíba com base na teoria da entropia. *Agriambi* **7**: 269-274.
- [22] Silva VPR. 2004. On climate variability in Northeast Brazil. *Journal of Arid Environment* **58**: 575-596.
- [23] Silva VPR, Sousa FSS, Cavalcanti, EP, Souza EP, Silva BB. Teleconnections between sea-surface temperature anomalies and air temperature. *Journal of Atmospheric and Solar-Terrestrial Physics* **68**: 781-792.
- [24] Singh VP. 1997. The use of entropy in hydrology and water resources. *Hydrological Processes* **11**: 587-626.
- [25] Singh VP, Rajagopal AK. 1987. Some recent advances in application of the principle of maximum entropy (POME) in hydrology. *IAHS-AISH Publication* **64**: 353-364.
- [26] Soofi E. 1997. Information theoretic regression methods. In: Fomby, T., Carter Hill, R. (Eds.), *Adv Econometrics – Applying Maximum Entropy to Econometric Problems* 12.



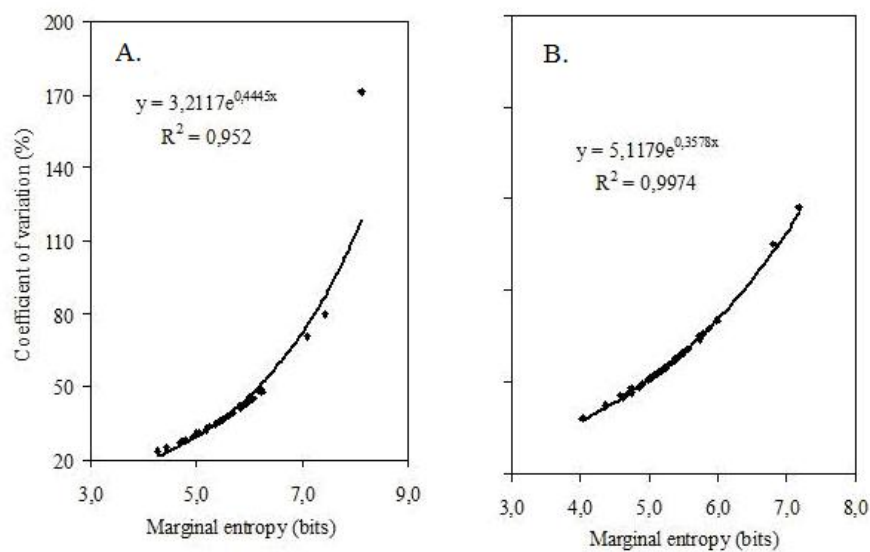


Fig.1

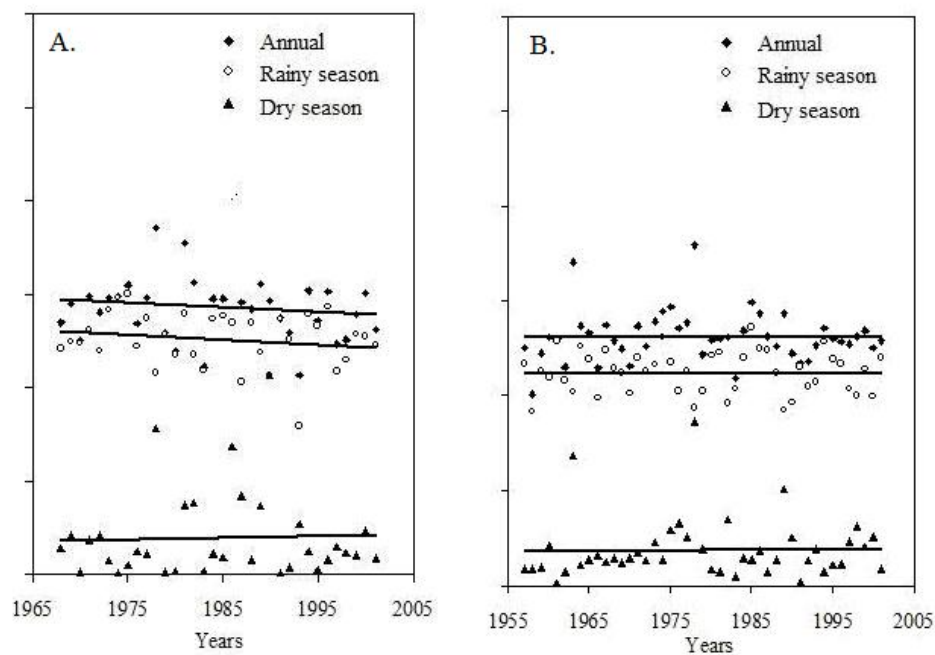


Fig.2

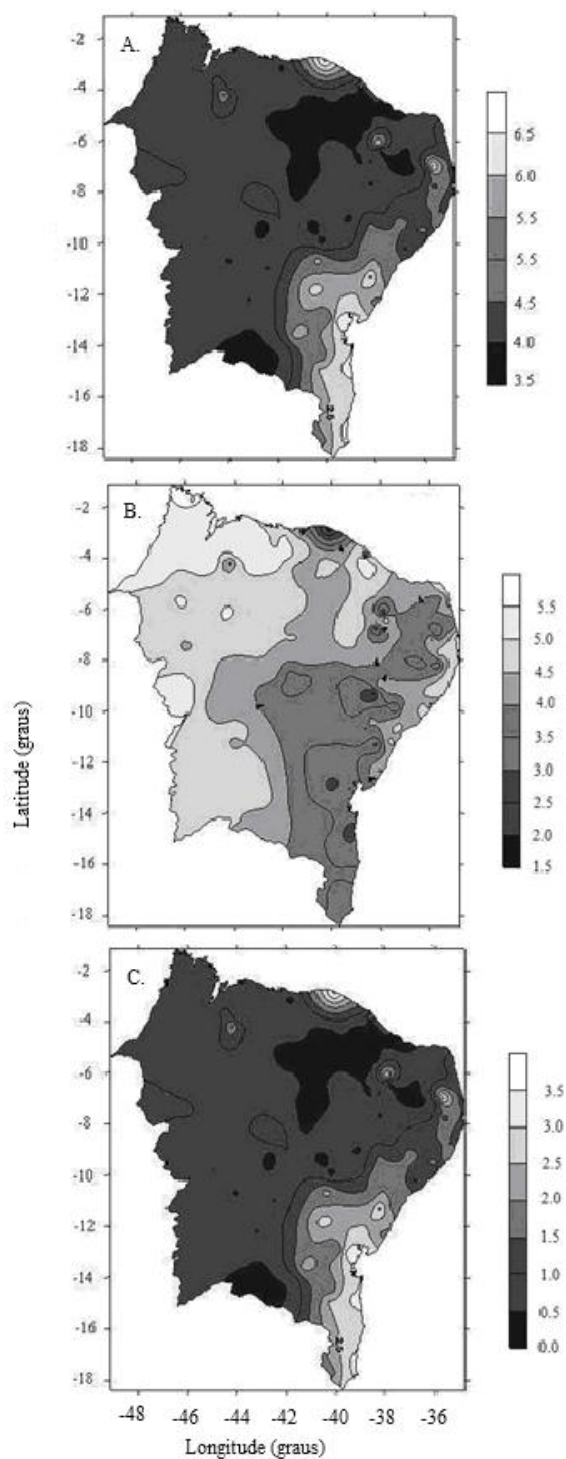


Fig.3

Table.1

	Icó station		
	Annual	Rainy	Dry
Marginal entropy bits)	5.2	4.5	0.8
CV (%)	34.2	44.9	39.2

São Luiz do Curu station			
Marginal entropy bits)	5.7	5.0	0.7
CV (%)	44.4	114.7	84.4

Table.2

Variables	Annual		Rainy season		Dry season	
	Trend	p-value	Trend	p-value	Trend	p-value
Icó station						
Rainfall (R)	-0.84	0.952	-0.502	0.741	0.4429	0.332
Marginal entropy in R	0.004	0.787	0.0013	0.667	0.0025	0.204
São Luiz do Curu station						
Rainfall (R)	-6.60	0.496	-5.67	0.447	-0.21	0.920
Marginal entropy in R	0.04	0.003	0.011	0.733	0.034	0.001